Abstract
This paper addresses the problem of bilingual lexicon extraction and lexical transfer selection, in the framework of computer-aided and machine translation. The method relies on parallel corpora, annotated at part of speech and lemma level. We first extract a bilingual lexicon using unsupervised statistical techniques. For each word with more than one translation candidates we build context vectors, based on the annotated parallel corpus information, in order to aid the selection of the contextually correct translation equivalent. The method achieves an overall precision of ca. 85% while the maximum recall reaches 75%.

Keywords: parallel corpora, bilingual lexicon extraction, translational ambiguity, translation prediction, lexical transfer selection

1 Introduction
The emergence of parallel corpora has evoked the appearance of many methods that attempt to deal with different aspects of computational linguistics (Véronis 00). Of special significance in the field of lexicography, terminology and machine translation is the impact of “bitexts”: a pair of texts in two languages, where each text is a translation of the other (Melamed 97). Such texts are necessary for providing evidences of use, directly deployable in statistical-based methodologies and enhance the automatic elicitation of the otherwise sparse linguistic resources.

This paper describes the design and development of a method for automatic bilingual lexicon extraction from a parallel bilingual corpus. Of particular importance is the integration of a lexical transfer selection strategy, which enables the rendering of the contextually correct translation for a given word. In this framework, we explore the relationship between word-senses and word-uses in a bilingual environment. We also analyse the way they can be represented in a context vector model for word translation prediction.

In the next section we give a brief overview of previous work in the field of automatic lexicon extraction from parallel corpora. In section 3 we present the proposed method, which aims at formulating and applying a context vector representation towards a context-based solution of translational ambiguities. In section 4 we analyse the evaluation process and the current results, while in sections 5 and 6 we discuss the conclusions drawn as well as possible applications and future enhancements.

2 Background
Recent developments in computer-aided and machine translation have moved towards the use of parallel corpora, aiming at two primary objectives: (i) to overcome the sparseness of the necessary resources and (ii) to avoid the burden of producing them manually. Furthermore, parallel corpora have proven rather useful for automatic dictionary extraction, which offers the advantages of a lexicon capturing the corpus specific translational equivalences, as Brown has pointed in (Brown 97; Piperidis et al. 00). Extending this approach, we investigate the possibility to use bilingual corpora in order to extract translational correspondences coupled with information about the wordsenses and contextual use. In particular, we focus on polysemous words with multiple translational equivalences.

The relation between word and word-usage, as compared to the relation of word and word-sense has been thoroughly addressed (Gale et al. 92a; Yarowsky 93; Kilgarriff 97). In this scope, we argue that through the exploitation of parallel corpora and without other external linguistic resources, we can adequately resolve the task of target word selection in computer-aided and machine translation. Along this line, it has been argued that the accumulative information added by a second language could be very important in lexical ambiguity resolution in the first language (Dagan et al. 91), while tools have been implemented for translation prediction, by using context information extracted from a parallel corpus (Tiedemann 01).

In addition, research on word sense disambiguation is significant in the design of a methodology for automatic lexical transfer selection. Although monolingual word sense disambiguation and translation are
perceived as different problems (Gale et al. 93), we examine whether certain conclusions, which are extracted during the process of word sense disambiguation, could be useful for translation prediction.

The role of context, as the only means to identify the meaning of a polysemous word (Ide & Véronis 98), is of primary importance in various statistical approaches. Brown in (Brown et al. 91) and Gale in (Gale et al. 92b; Gale et al. 93), use both the context of a polysemous word and the information extracted from bilingual aligned texts, in order to assign the correct sense to the word. Yarowsky explores the significance of context in creating clusters of senses (Yarowsky 95), while Schütze in (Shütze 98), addresses the sub-problem of word sense discrimination through the context-based creation of three cascading types of vectors.

We examine the impact of context upon translation equivalent selection, through an “inverted” word sense discrimination experiment. Given the possible translational candidates, which are extracted from the statistical lexicon, we investigate the discriminant capacity of the context vectors, which we build separately for each of the senses of the polysemous word.

3 Proposed method

The basic idea underlying the proposed method is the use of context vectors, for each of the word usages of a polysemous word, in order to resolve the problem of lexical transfer selection. The method consists of three stages:

- Bilingual Lexicon Extraction
- Context Vectors Creation
- Lexical Transfer Selection

The first stage could be omitted, if a bilingual lexicon is already available. Figure 1 depicts an overall view of the system’s architecture.

3.1 Lexicon Building

The first goal is to build a bilingual lexicon. Parallel corpora are sentence-aligned using a Gale & Church-like algorithm (Gale et al. 91) and annotated on both language sides for part-of-speech and lemma. Focusing on the semantic load bearing words, we filter the pos tagged corpus and retain only nouns, adjectives and verbs. The corpus-specific lexicon is extracted using unsupervised statistical methods, based on two basic principles:

- No language-pair specific assumptions are made about the correspondences between grammatical categories. In this way, all possible correspondence combinations are produced, as this is possible during the translation process from a source language to the target language.

- For each aligned sentence-pair, each word of the target sentence is a candidate translation for each word of the aligned source sentence.

Following the above principles we compute: the absolute frequency (the number of occurrences) of each word and the frequency of each word pair, in a sentence pair. Using these frequencies, we extract lexical equivalences, based on the following criteria:

1. The frequency of the word pair must be greater than threshold $Thr_1$.

2. Each of the conditional probabilities $P(W_t|W_s)$ and $P(W_s|W_t)$ has to be greater than threshold $Thr_2$.

3. The product $P(W_t|W_s) \cdot P(W_s|W_t)$ must be greater than threshold $Thr_3$. This product is indeed the score of the translation.

After experimentation and examination of the results, $\{Thr_1, Thr_2, Thr_3\}$ were set to $\{5, 0.25, 0.15\}$. More specifically, experiments with $Thr_1=1$
and $Thr_1=10$ revealed that the former resulted in high recall with significantly low precision, i.e. a fairly large but low quality lexicon, while the latter resulted in relatively high precision, whilst recall was radically reduced, i.e. a fairly small high-quality lexicon. We empirically decided to fix $Thr_1$ in the middle of the experimentation range. $Thr_2$ was set to 0.25 to account for a maximum number of 4 possible translations, which a presumed polysemous word could have, according to the corpus data. $Thr_3$ was set to 0.15 to account for the lower bound of the product of conditional probabilities $P(W_i|W_x)$ and $P(W_x|W_i)$, taking into consideration the threshold of the individual conditional probabilities, i.e. we empirically set the lower bounds to the conditional probabilities as $\{0.25, 0.6\}$.

The extracted bilingual lexical equivalences account for: words with one translation (90% of total) and words with multiple translations (10%). Words with multiple extracted equivalents are further distinguished in: (i) words that have multiple, different in sense, translational equivalents, (ii) words that have multiple, synonymous translations and (iii) words that have multiple translations, which are in fact wrong due to statistical errors of the method. In the following, we focus on (i), that is polysemous words, with multiple translational equivalents, which cannot mutually replace each other in the same context.

### 3.2 Context Vectors Creation

For each word, in the set of the treated grammatical categories, in the corpus, a context vector is created based on: (i) the extracted lexicon, in order to retrieve the possible translations of a word and (ii) the parallel aligned sentences to retrieve those words that systematically co-occur with that word (thus contributing to the definition of its meaning). The process is described below:

**Step 1.1:** For “univocal” words, words with only one translation, we assume that the translation is also its “sense”. For the untranslated words we make no assumption, though they also participate in the created vectors. Both these categories of words are denoted by $W_u$.

**Step 1.2:** For the words with multiple translations in the extracted lexicon, we cannot automatically pick out the polysemous words. Therefore for each of these $W_p$, we make the following assumption: *Each word, which has more than one translation in the lexicon, could potentially be a polysemous one.* We suppose $W_p$ is one of those words and let $T_1$ and $T_2$ be two possible translations. When $W_p$ is found in a source sentence, we search for the words $T_1$ and $T_2$ in the target sentence. If one and only one is matched, e.g. $T_1$, we conclude that this is the correct translation and $W_p$ is replaced in the source sentence by $W_p.T_1$. This is repeated for each word with multiple translations. In the case of erroneous multiple translations (caused by statistical errors), none of $T_1$ or $T_2$ are assigned as a sense, due to the simultaneous appearance of more than one in the target sentence. In the end, we have a new corpus in which some words appear as before and some have been labeled by their “local senses”.

**Step 2:** In order to build the context vectors, we address the words $W_p.T_1$ and $W_p.T_2$ as being different. Then we isolate those source sentences where either “word-sense” $W_p.T_1$ or $W_p.T_2$ appear exclusively. In each different set of sentences we examine the context of $W_p.T_i$ in a window of certain length centered to the word of interest $W_p.T_i$. The size of the window is defined as:

$$\text{window\_size} = 2n,$$

where $n$ denotes the number of word tokens on either side of the word of interest.

Inside this window of token-words, we look for words, $W_x$, which belong to the selected grammatical categories. Each of $W_x$ is added to the vector, along with the number of times this word has appeared in the context of the word $W_p.T_i$. We follow a similar procedure for each word $W_u$.

**Step 3:** The final formation of the context vectors is based on the following equations:

$$N_{W_x.W_p.T_i} \geq k$$

$$P(W_x|W_p.T_i) \geq \alpha_1,$$

where $N_{W_x.W_p.T_i}$ is the number of the total co-occurrences of words $W_x$ in the window of $W_p.T_i$, $k$ is the minimum co-occurrences that a word $W_x$ must have in order to participate in the context vector which describes the word $W_p.T_i$, $P(W_x|W_p.T_i)$ is the conditional probability of the word $W_x$ given the appearance of $W_p.T_i$ and $\alpha_1$ is a threshold, which the probability $P(W_x|W_p.T_i)$ must exceed. $P(W_x|W_p.T_i)$ is also the score of $W_x$ in the context vector of $W_p.T_i$.

In the case of a word $W_u$, with only one translation or no translation in the lexicon, similar equations are used:

$$N_{W_x.W_u} \geq k$$

$$P(W_x|W_u) \geq \alpha_2,$$
where \( N_{W,W_x} \) is the number of the total co-occurrences of words \( W_x \) in the window of \( W_u \), \( k \) is defined as the minimum co-occurrences of \( W_x \) and \( W_u \), \( P(W_x|W_u) \) is the conditional probability of the word \( W_x \) given the appearance of \( W_u \) and \( a_2 \) is a threshold, which the probability \( P(W_x|W_u) \) must exceed. \( P(W_x|W_u) \) is the score of \( W_x \) in the context vector of \( W_u \).

Whenever at least one of these criteria in each set of equations is not met, the word \( W_x \) is deleted from the vector. In (3) and (5) we use two distinct thresholds \( a_1 \) and \( a_2 \) with

\[
a_1 < a_2, \quad (6)
\]
as we would like polysemous words to have greater vectors. The parameters \( a_1 \) and \( a_2 \) are set, after experimentation, to 0.05 and 0.1 respectively.

Thus, we have constructed, a context vector for each word in the set of the specified grammatical categories. The vector consists of words that systematically co-occur with the word of interest and their context-vector scores. In Figure 2 we present the flow diagram for the creation of context vectors. For each new word the process is iterated over the words of the corpus.

### 3.3 Lexical Transfer Selection

Based on the lexicon and the context vectors, the algorithm can disambiguate an ambiguous word, when appearing in a certain context, by comparing this context with the previously created vectors of its “senses”. The process includes the following steps:

**Step I:** Let \( W_p \) be an ambiguous word, with \( T_i \) translational equivalents extracted from the lexicon. A sentence is fed to the system and \( W_p \) is one of the words. Each \( T_i \) of the “senses” \( W_p,T_i \) are considered to be translation candidates for the sentence at hand.

**Step II:** For each of \( W_p,T_i \) an extended vector \( V_{xyzw} \) is produced. The main characteristic of the expanded vector is its depth \( d_i \). The depth \( d_i \) denotes the number of “co-occurrence connections” between words, which we use in order to “meaningfully connect” the word \( W_p,T_i \) with any word \( W_x \). In our methodology:

\[
d_i = 4 \quad (7)
\]

We believe that a greater value for \( d_i \) would capture the spurious co-occurrences of words, thus it would not represent a logical and linguistically expected “sense-connectivity”. The vector of \( W_p,T_i \) consists of the words \( W_x \) that appear in the context vector \( V_1 \) of \( W_p,T_i \), the words that appear in the context vectors \( V_i \) of each word in \( V_1 \), and so on until depth \( i \) (\( V_i \) words are in depth \( i \), \( V_{i+1} \) words are in depth \( i+1 \) etc). Figure 3 in the next page shows a diagram for the created vector.

**Step III:** Each of the word \( W_x \) that participates in the enlarged vector is assigned an extended-vector score \( EVScore_{xw,p} \) or \( EVScore_{xw,u} \), depending on

![Figure 2: Context-Vector Creation Procedure](image-url)
the type of the word with which it co-occurs:

\[ EV\text{Score}_{x|W_p,T_i} = \frac{P(W_x|W_p,T_i)}{2^{1-d_x}} \]  

\[ EV\text{Score}_{x|W_u} = \frac{P(W_x|W_u)}{2^{1-d_x}}, \]  

where \( P(W_x|W_p,T_i) \) and \( P(W_x|W_u) \) are defined in (3) and (5) and \( d_x \) is the depth in which \( W_x \) was found. In case of multiple appearances of \( W_x \) in the extended vector, we choose the one in the lowest depth, as it is the most significant in the process of defining the sense of \( W_p \).

**Step IV:** The final lexical transfer selection procedure examines each extended vector of \( W_p,T_i \) separately. We compare the words inside the \( \pm n \) window of word \( W_p \) of the sentence under examination with those included in the vector. For each matched word we compute the appropriate score, using (8) and (9). By adding the scores of the matched words, we assign to each possible translational equivalent \( T_i \) a total score, depending on the associated extended vector. Finally, for the lexical transfer selection, we choose the word-sense \( W_p,T_i \) with the highest score. If both scores are equal, the algorithm does not choose randomly and can output both as candidate translations. A feedback mechanism could be foreseen to minimize these cases, if appropriate, in a subsequent transfer selection round.

## 4 Results - Evaluation

The corpus used was the INTERA parallel corpus (Gavrilidou et al. 04) consisting of official EU documents in English and Greek from five different domains; education, environment, health, law and tourism. The corpus comprises 100,000 aligned sentences, containing on average 830,000 tokens of the selected grammatical categories (nouns, verbs and adjectives) in either language. The corresponding lemmas are 20,000. The complete bilingual lexicon comprises 5280 records (where multiple translational equivalences of a word are counted as one record).

Evaluation was focused on the ability to resolve truly ambiguous words, leaving aside words with synonymous translations, or erroneous translational candidates. For this purpose, a set of ambiguous words in English, the contextually correct translation equivalent of which is “univocal” in Greek, were manually selected. The selected set was \{active, floor, seal, settlement, solution, square, vision\}.

For the above words, we extracted the sentences, in the parallel corpus, that contain them. We adopted the 10-fold cross validation technique for evaluation, computing the average results over the 10 iterations of the algorithm. The possible answers, given by the algorithm were:

- **Correct**, when only the selected translational equivalent was present in the target sentence.
- **Wrong**, when the selected translational equivalent was different from the one appearing in the target sentence.
- **No answer**, when the translational equivalents were assigned the same score.

Precision was calculated as the ratio of the correct answers to the sum of correct and wrong answers. Recall was calculated as the ratio of the correct answers to the possible correct answers. The experiment was first performed with three sets of \( k, n \): (i) \( k=3 \) and a window of \( n=5 \), (ii) \( k=3 \) and a window of \( n=7 \), and (iii) \( k=3 \) and a window of \( n=15 \) (\( k \) referring to (2) and (4)). The averaged results over the 10 iterations are shown in Table 1.
In order to simulate a larger corpus, we enlarged the produced context vectors. We conducted the experiment again, with \( k=1 \) and the three variants for the size of the windows defined as previously. The results are shown in Table 2.

<table>
<thead>
<tr>
<th>( k = 3 )</th>
<th>( n = \pm 5 )</th>
<th>( k = 3 )</th>
<th>( n = \pm 7 )</th>
<th>( k = 3 )</th>
<th>( n = \pm 15 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>85.5</td>
<td>90.8</td>
<td>91.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong</td>
<td>9.6</td>
<td>14.5</td>
<td>26.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No answer</td>
<td>25.1</td>
<td>15.2</td>
<td>4.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>89.9%</td>
<td>86.2%</td>
<td>77.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>71.1%</td>
<td>75.3%</td>
<td>76.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Answered</td>
<td>79.1%</td>
<td>87.4%</td>
<td>98.0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Results for \( k=3 \)

<table>
<thead>
<tr>
<th>( k = 1 )</th>
<th>( n = \pm 5 )</th>
<th>( k = 1 )</th>
<th>( n = \pm 7 )</th>
<th>( k = 1 )</th>
<th>( n = \pm 15 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>88.1</td>
<td>92.9</td>
<td>87.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong</td>
<td>11.7</td>
<td>17.0</td>
<td>31.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No answer</td>
<td>20.7</td>
<td>10.6</td>
<td>1.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>88.2%</td>
<td>84.5%</td>
<td>73.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>73.1%</td>
<td>77.1%</td>
<td>72.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Answered</td>
<td>82.8%</td>
<td>91.2%</td>
<td>98.8%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Results for \( k=1 \)

In Figure 4 we present the created vector \( V_1 \) for the polysemous word “solution” and in Figures 5 and 6 we show two examples of our system’s behavior.

As expected, the wider the window, the more likely that the system gives an answer, although the precision decreases. Especially for a window size \( n=15 \), which in most cases in the given corpus contains all the tokens in a sentence, we notice that the system’s performance declines disproportionately. This is due to multiple erroneous statistical co-occurrences that are semantically irrelevant is such wide windows.

In the second experiment, the percentage of the answered cases and recall increased, compared to the first experiment, while precision slightly decreased. The results also indicate that although a smaller window and a higher absolute appearance threshold \( k \) would lead to a lower number of answers, the accuracy increases.

To evaluate the performance of the method taking into consideration the special characteristics of our corpus, we computed a “baseline” performance (Gale et al. 92c). We assign to each polysemous word, found in the test set, the most frequent of its possible senses. The estimated baseline performance was 55% on average due to the almost equal distribution of the different senses of the words. Thus, the employment of context vectors method lead to an increase in recall of almost 20%.

![Figure 4: First level vector for “solution”](image)

<table>
<thead>
<tr>
<th>Polysemous Word ( W_p )</th>
<th>Words in First Level Vector ( V_1 ) of ( W_p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>agenase ( _T )</td>
<td>concentrate ( _T )</td>
</tr>
<tr>
<td>aqueous ( _T )</td>
<td>contain ( _T )</td>
</tr>
<tr>
<td>capsule ( _T )</td>
<td>contraindicated ( _T )</td>
</tr>
<tr>
<td>child ( _T )</td>
<td>fill ( _T )</td>
</tr>
<tr>
<td>clear ( _T )</td>
<td>insulin ( _T )</td>
</tr>
<tr>
<td>colourless ( _T )</td>
<td>inject ( _T )</td>
</tr>
<tr>
<td>homogeneous liquid ( _T )</td>
<td>patients ( _T )</td>
</tr>
<tr>
<td>problem ( _T )</td>
<td>pen ( _T )</td>
</tr>
<tr>
<td>possible ( _T )</td>
<td>vial ( _T )</td>
</tr>
</tbody>
</table>

In Figure 5 we present the created vector \( V_1 \) for the polysemous word “solution” and in Figures 5 and 6 we show two examples of our system’s behavior.

![Figure 5: First example of Lexical Transfer Selection](image)

<table>
<thead>
<tr>
<th>Test Sentence</th>
<th>Polysenous word ( W_p = solution )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W_p _T_1 = solution _διάλυµα ) (solution, as a homogeneous liquid)</td>
<td>( W_p _T_2 = solution _λύση ) (solution, as answer, decision)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context Vectors Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W_p _T_1 = solution _διάλυµα ) be: score=0.3 clear: score=0.11 colourless: score=0.08 visible: score=0.025 use: score=0.0475</td>
</tr>
<tr>
<td>( W_p _T_2 = solution _λύση ) be: score=0.44</td>
</tr>
<tr>
<td>Correct Selected: ( W_p _T_1 = solution _διάλυµα )</td>
</tr>
</tbody>
</table>

![Figure 6: Second example of Lexical Transfer Selection](image)
5 Applications and Future work

The proposed method can be used as a translational tool, for computer-aided and machine translation, especially as it concerns translation customization processes. Furthermore the method can be used as feedback mechanism for a refinement in statistical lexicon extraction. As a validation, we conducted a second experiment, over all the words in the lexicon, which had multiple translations (although not always correct). The results were similar to the ones presented in Table 1 and Table 2. Thus, such methods can be of utmost importance for bootstrapping the development of multilingual lexica with semantic constraints on the potential cross-lingual equivalences. Forthcoming experiments will include tests on larger corpora and use of linguistically principled window selection.

6 Conclusions

We presented a statistical method for lexical transfer selection, with special attention to polysemous words. Our technique relies on a bilingual parallel, aligned and annotated corpus, without resorting to other external linguistic resource. The method is language independent and suitable for translation prediction for any language pair.

Based on the aligned sentences, we extracted a statistical bilingual lexicon, from which we identified words with multiple translational equivalents. We then extracted context vectors, representing the impact of adjacent words to the sense of an ambiguous word. Finally, we merge the information derived from the context and the lexicon to obtain the selection of the contextually correct translational equivalent. Evaluation shows a promising overall performance, as compared to the evaluated baseline performance.

References


[Dagan et al. 91]: Ido Dagan, Alon Itai, Ulrike Schwall, Two languages are more informative than one, Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California, June 1991, pp. 130-137.


[Gale et al. 92a]: William A. Gale, Kenneth W. Church and David Yarowsky One sense per discourse, Proceedings of the Speech and Natural Language Workshop, San Francisco, Morgan Kaufmann, pp. 253-237.


[Piperidis et al. 00]: Stelios Piperidis, Harris Papageorgiou and Sotiris Boutsis From sentences to words and clauses, Parallel Text Processing, Jean Véronis (editor), Kluwer Academic Publishers, pp. 117-138.


[Tiedemann 01]: Jörg Tiedemann, Predicting Translations in Context, Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP), Tzigov Chark, Bulgaria, September 2001, pp. 240-244.
